+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 26/02/2026

S'inscrire

Formation IA Souveraine : Axolotl, Kubernetes & Llama 3

3 jours (21 heures)

Présentation

Déployez une IA souveraine en interne en combinant Axolotl (fine-tuning), Kubernetes (industrialisation) et Llama 3 (LLM open weights). Vous apprendrez à entraîner, servir et sécuriser des modèles pour des cas d'usage concrets : assistants métiers, RAG sur documents, classification et extraction.

Cette formation vise à rendre opérationnelle une chaîne complète : préparation des données, adaptation de Llama 3 via Axolotl (LoRA/QLoRA), packaging des artefacts et mise en production sur Kubernetes avec observabilité et contrôle des accès.

Notre approche est 100% pratique : ateliers guidés, démos reproductibles, scripts d'automatisation. Livrables : repo de configuration Axolotl, manifests Kubernetes (Deployment/Service/Ingress), pipeline de build, checklist sécurité et runbook d'exploitation afin de valider vos acquis.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Préparer et valider un dataset d'entraînement adapté au cas d'usage.
- Fine-tuner Llama 3 avec Axolotl (LoRA/QLoRA) et évaluer les résultats.
- Servir le modèle (vLLM/TGI) et gérer la montée en charge.
- Déployer sur Kubernetes avec GPU, autoscaling et stockage.
- Sécuriser et superviser l'inférence (auth, logs, métriques, coûts).

Public visé

- Ingénieurs ML / MLOps
- DevOps / SRE
- Architectes cloud / plateforme
- Développeurs backend intégrant des LLM

Pré-requis

- Bonnes bases Linux et ligne de commande
- Notions Python (environnements, dépendances)
- Fondamentaux Docker et images
- Connaissances Kubernetes (pods, services, ingress)
- Notions de base en NLP/LLM (tokenisation, prompts)

Pré-requis techniques

- 32 Go RAM recommandés (16 Go minimum), CPU 8 cœurs
- GPU NVIDIA recommandé : 16 Go VRAM minimum (24 Go+ idéal), drivers + CUDA
- OS : Linux ou macOS ; Windows via WSL2 possible
- Outils : Docker, kubectl, Helm, Git, éditeur de code
- Accès à un cluster Kubernetes (local ou distant) avec nœuds GPU si entraînement/inférence accélérés

Programme de formation IA Souveraine : Axolotl, Kubernetes & Llama 3

[Jour 1 - Matin]

Fondations d’une IA souveraine et prise en main de Llama 3

- Définir les objectifs de souveraineté : données, modèles, infrastructure, traçabilité
- Panorama Llama 3 : tailles, licences, contraintes matérielles, cas d’usage (chat, RAG, classification)
- Choisir un format de déploiement : weights, quantification (4/8 bits), CPU vs GPU
- Mettre en place un environnement reproductible : Python, CUDA, drivers, gestion des versions
- Atelier pratique : Lancer une inférence locale Llama 3 et mesurer latence/VRAM.

[Jour 1 - Après-midi]

Préparer l’adaptation du modèle avec Axolotl (LoRA/QLoRA)

- Comprendre l’approche fine-tuning vs instruction tuning vs RAG (critères de choix)
- Structurer un dataset : formats (JSONL), champs (instruction/input/output), règles de qualité

- Configurer Axolotl : modèle de base, tokenizer, hyperparamètres, batch/grad accumulation
- Mettre en place l'évaluation : jeux de tests, métriques simples, tests de non-régression
- Atelier pratique : Construire un dataset d'instructions et générer une config Axolotl prête à entraîner.

[Jour 2 - Matin]

Entraîner et versionner un adaptateur LoRA avec Axolotl

- Lancer un entraînement LoRA/QLoRA : paramètres clés (lr, epochs, warmup, cutoff_len)
- Optimiser les coûts : gradient checkpointing, quantification, choix du rank/alpha LoRA
- Suivre l'entraînement : logs, courbes, détection d'overfitting, early stopping
- Versionner les artefacts : adaptateurs, config, dataset, reproductibilité (seed, commit)
- Atelier pratique : Exécuter un fine-tuning QLoRA, exporter l'adaptateur et valider sur un set de tests.

[Jour 2 - Après-midi]

Packaging et serving du modèle pour Kubernetes

- Choisir un serveur d'inférence : vLLM / TGI / llama.cpp (critères : perf, GPU, streaming)
- Construire une image container : dépendances, cache des weights, démarrage déterministe
- Exposer une API : endpoints, timeouts, streaming, limites de tokens, concurrency
- Gérer la configuration : variables d'environnement, fichiers, séparation modèle/adaptateur
- Atelier pratique : Conteneuriser un serveur d'inférence Llama 3 + LoRA et tester l'API en local.

[Jour 3 - Matin]

Déployer sur Kubernetes : GPU, scaling et fiabilité

- Déployer avec manifests : Deployment/StatefulSet, Services, probes (liveness/readiness)
- Activer l'accès GPU : device plugin, requests/limits, node selectors, taints/tolerations
- Gérer la montée en charge : HPA, autoscaling basé sur métriques, files d'attente
- Observabilité : logs structurés, métriques (latence, tokens/s), alerting
- Atelier pratique : Déployer l'inférence sur un cluster Kubernetes et valider la stabilité via probes et charge.

[Jour 3 - Après-midi]

Sécurité, conformité et exploitation d'une IA souveraine

- Sécuriser l’accès : authn/authz, réseau (NetworkPolicies), quotas, rate limiting
- Gérer secrets et données : Secrets, chiffrement, politiques de rétention, traçabilité des prompts
- Gouvernance du modèle : registre interne, validation, rollback, contrôle des versions
- Bonnes pratiques d’exploitation : runbooks, tests de dérive, monitoring qualité des réponses
- Atelier pratique : Mettre en place une politique d’accès + journalisation minimale et un plan de rollback du modèle.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.